Trabajo práctico 2

Sintetización de fonemas mediante un modelo AR

1. Introducción

Un enfoque para producir la síntesis de voz artificial es imitar el proceso de generación del habla humana. Considerando el modelo de resonancia de la Figura 1, la producción de sonidos de habla se puede pensar como una fuente de señal que excita un sistema que resonará más en ciertas frecuencias que en otras, dependiendo la configuración que tenga el tracto vocal para cada fonema particular [1]. Uno de los métodos más conocidos para modelas este problema se basa en la técnica LPC (Linear Predictive Coding), con la que se podrá modelar la producción de habla a través de un sistema LTI excitado por un proceso aleatorio. No obstante, el objetivo de este trabajo práctico consiste solo en sintetizar realizaciones de un proceso aleatorio que se modela mediante un sistema lineal, pero no la estimación de los parámetros que lo definen (que serán datos conocidos en nuestro caso).

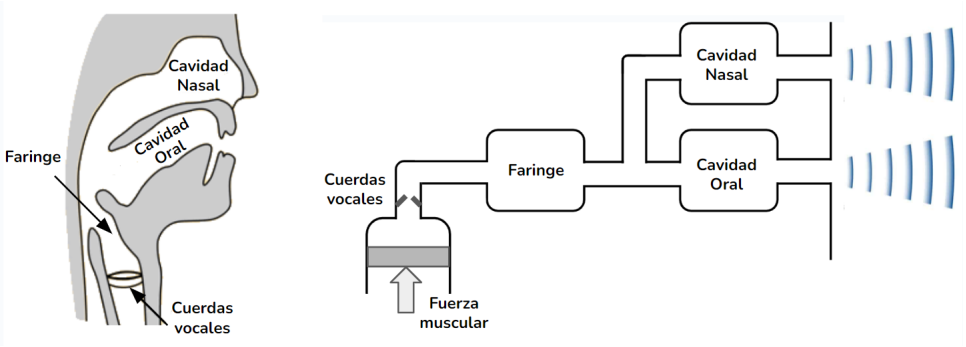

Figura 1: Modelo de resonancia del sistema de generación de habla. A la izquierda modelo anatómico. A la derecha, un modelo idealizado.

1.1. Descripción del modelo

El modelo utilizado por LPC para la generación habla, como se muestra en la Figura 2, es representado con un sistema LTI de tipo IIR (Infinite Impulse Response) con el que se produce un proceso autoregresivo (AR) de orden P (cantidad de coeficientes) [2]. Este modelo modifica el espectro del proceso de entrada para hacer resonar diferentes bandas de frecuencia, emulando de esta manera la resonancia del tracto vocal de un determinado fonema. Cabe destacar que una señal acústica (como al voz humana) es un proceso no estacionario, pero que puede asumirse estacionario para tramos de audio de corta duración.

Para modelar un fonema, existen dos clases de sonidos dependiendo la fuente de excitación: de tipo *sonora*, cuando el fonema posee una exitación periódica (que emula el sonido de las cuerdas vocales), o de tipo *sorda*, cuando la excitación es un proceso blanco (no actuan las cuerdas vocales). Por ejemplo, las vocales (“a”, “e”, “i”, “o”, “u”) necesitan una excitación

sonora. Otros fonemas se pueden producir con una excitación sorda, como es el caso de las consonantes fricativas ("s", "f", "j", "sh"). Por otro lado, algunas consonantes poseen una combinación mas compleja, por lo que en este trabajo solo nos centraremos en los casos mencionados anteriormente. Para el habla sonora, la fuente de excitación más apropiada es un tren de impulsos periódico[1] $U(n) = \sqrt{T f_s} \sum_{k=-\infty}^{\infty} \delta(n - kT f_s)$ de periodo T, siendo $f_0 = 1/T$ el "*pitch*" o frecuencia fundamental de la excitación. En esta última expresión, f_s representa la frecuencia de muestreo con la que se digitalizaron las señales analógicas que forman parte del análisis. Por su parte, el habla sorda recibe como entrada un proceso de ruido blanco gaussiano $U(n) \sim N(0, 1)$, que resulta más adecuado para este caso. El sistema LTI-IIR utilizado para modelar este proceso posee coeficientes $\{a_1, a_2, ..., a_P\}$ en la realimentación de la ecuación en diferencias y una ganancia b [2], cuya entrada es el proceso de excitación $U(n)$ y su salida $X(n)$, como se expresa en la ecuación (1).

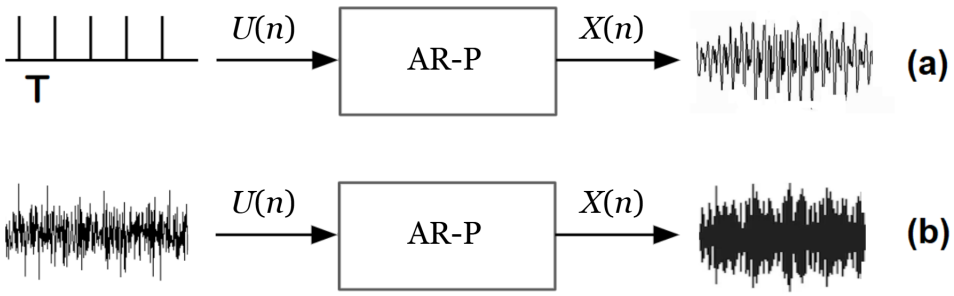

Figura 2: Modelos de generación del habla mediante un sistema LTI-IIR de orden P. (a) señales sonoras (la excitación es un tren de impulsos periódico de frecuencia $f_0 = 1/T$). (b) señales sordas (la excitación es ruido blanco).

$$X(n) = \sum_{i=1}^{P} a_i X(n - i) + b \cdot U(n) \tag{1}$$

Tomando como referencia este modelo, se puede hallar la PSD $S_X(\omega)$ a la salida del sistema en función de la PSD del proceso de entrada $S_U(\omega)$ de acuerdo a la siguiente relación:

$$S_X(\omega) = |H(\omega)|^2 S_U(\omega) = \frac{b^2}{\left|1 - \sum_{i=1}^{P} a_i \exp\left(-j\omega\, i\right)\right|^2} S_U(\omega), \tag{2}$$

donde $H(\omega)$ es la respuesta en frecuencia del sistema LTI de la ecuación (1). Alternativamente, dado que nos interesa ver el comportamiento en frecuencia de una señal analógica que fue digitalizada, se puede definir una expresión de la PSD en función de la frecuencia en Hz, mediante el cambio de variable $\omega = 2\pi f / f_s$:

$$S_X(f) = \frac{b^2}{\left|1 - \sum_{i=1}^{P} a_i \exp\left(-j2\pi \frac{f}{f_s}\, i\right)\right|^2} S_U(f), \tag{3}$$

[1]Esta definición representa una realización del proceso, ya que el tren de impulsos puede considerarse como un proceso periódico de fase aleatoria, por lo que podemos definir la realización para una fase particular.

2. Desarrollo

En el campus, el archivo `tp3.zip` contiene varias pistas de audio con la pronunciación de diferentes fonemas. Por un lado, las *vocales* "`a.wav`", "`e.wav`", "`i.wav`", "`o.wav`" y "`u.wav`", y las *consonantes fricativas* "`sh.wav`", "`f.wav`" , "`s.wav`", "`j.wav`". Los audios han sido muestreados a una frecuencia de $f_s = 14700\ Hz$ (asuma que son procesos aproximadamente estacionarios[2]). También se dispone del archivo "`data.py`" que contiene los diccionarios con los coeficientes "`coef_b`" y "`coef_a`" del proceso AR ($P = 20$) que modela cada uno de los audios provistos. Además se dispone de algunas funciones auxiliares.

Ejercicio 1: análisis del modelo

(a) A partir de los audios reales, grafique para cada fonema un tramo de 200 ms de señal. También estime y grafíque la función de autocorrelación en cada caso.

(b) Para cada audio calcule y grafique su periodograma (con su magnitud en decibeles[3] y en el rango de frecuencias $f \in [0, f_s/2)$), superpuesto a su respectiva PSD teórica de acuerdo a la ecuación (3). **Ayuda:** recuerde que $S_U(f) = 1$ para las consonantes fricativas, debido a que la excitación es un proceso blanco de varianza unitaria; por su parte, para el caso de las vocales, puede obtener $S_U(f)$ mediante la función `psd_pulsos()` definida en el archivo `data.py` (ver docstring).

Ejercicio 2; sintetización de fonemas

Se busca para cada fonema, sintetizar un proceso que emule el sonido de ese fonema a partir de los coeficientes del modelo propuesto. Considere una duración de 500 ms para cada señal generada. Adicionalmente, debe utilizarse la función `suavizar_bordes()`, disponible en `data.py` (ver docstring), para modular la amplitud de cada fonema, suavizando así los extremos de la señal. Para el caso de las vocales, como requieren un proceso periódico como excitación, se utilizará un tren de impulsos de frecuencia f_0 generado por la función `gen_pulsos()` definida en el archivo `data.py` (ver docstring).

(a) Sintetice cada fonema en base a los requerimientos mencionados, considerando para las vocales una frecuencia de pitch $f_0 = 200$ Hz. Suavice los bordes de las señales con un 20 % de transición y grafique los primeros 200 ms de cada fonema sintético.

(b) Genere una señal concatenando los nueve fonemas (las cinco vocales y las cuatro consonantes) obtenidos en el punto anterior. Reproduzca la secuencia resultante para evaluar subjetivamente los sonidos percibidos.

(c) Repita la consigna del punto anterior, pero generando los fonemas con un pitch diferente para cada vocal (Ej: $a\rightarrow$ 100 Hz, $e\rightarrow$ 125 Hz, $i\rightarrow$ 150 Hz, $o\rightarrow$ 125 Hz y $u\rightarrow$ 100 Hz).

3. Conclusiones

Como conclusiones, elabore un resumen breve y conciso comentando características que considere relevantes del método propuesto en este trabajo y los resultados obtenidos, así como dificultades encontradas (si fuera el caso) y cómo fueron abordadas.

[2]Si los audios no fueran estacionarios (en general no lo son) deberían analizarse por segmentos más cortos (30-100 ms) para garantizar la estacionariedad por tramos.

[3]PSD en decibeles: $S_X(\omega)|_{dB} = 10\log_{10}(S_X(\omega))$.

4. Normas y material entregable

- **Informe**: El informe debe entregarse en formato PDF (**no se aceptarán otros formatos**) y con nombre: `TP2_GXX.PDF` (donde `XX` es el número de grupo).No debe incluirse código fuente en el PDF. El código de Python debe presentarse en un archivo aparte.
- **Código**: El código debe incluirse junto al informe en un archivo ZIP (con mismo nombre que el informe) que deberá subirse al campus.
- Se recuerda a los estudiantes que las entregas deben ser un producto original de cada estudiante, por lo que se les pide revisar la sección 6 del programa de la materia y el Código de Honor y Ética de la Universidad.

Referencias

[1] M Olmo R Nave. Resonancia del Tracto Vocal. http://hyperphysics.phy-astr.gsu.edu/hbasees/Music/vocres.html

[2] Makhoul, J., "Linear Prediction: A Tutorial Review,ÏEEE Proceedings, Vol. 63, pp. 561-580, 1975.

[3] Stoica, Petre, and Randolph L. Moses , 2005. Spectral analysis of signals.